CSED 490C

LAB 3

Jeongseop Yi (49004543)

Q1

```
cnt = min(streamSize, inputSize - (numStreams - 1) * streamSize)
```

In each stream, there are `2 * cnt * sizeof(float)` bytes of data are moved from host to device. In total, there are `2 * inputLength * sizeof(float)` bytes of data are moved from host to device (input1, input2).

Q2

Pinned memory refers to an allocation of virtual memory on host for exclusive access by the GPU.

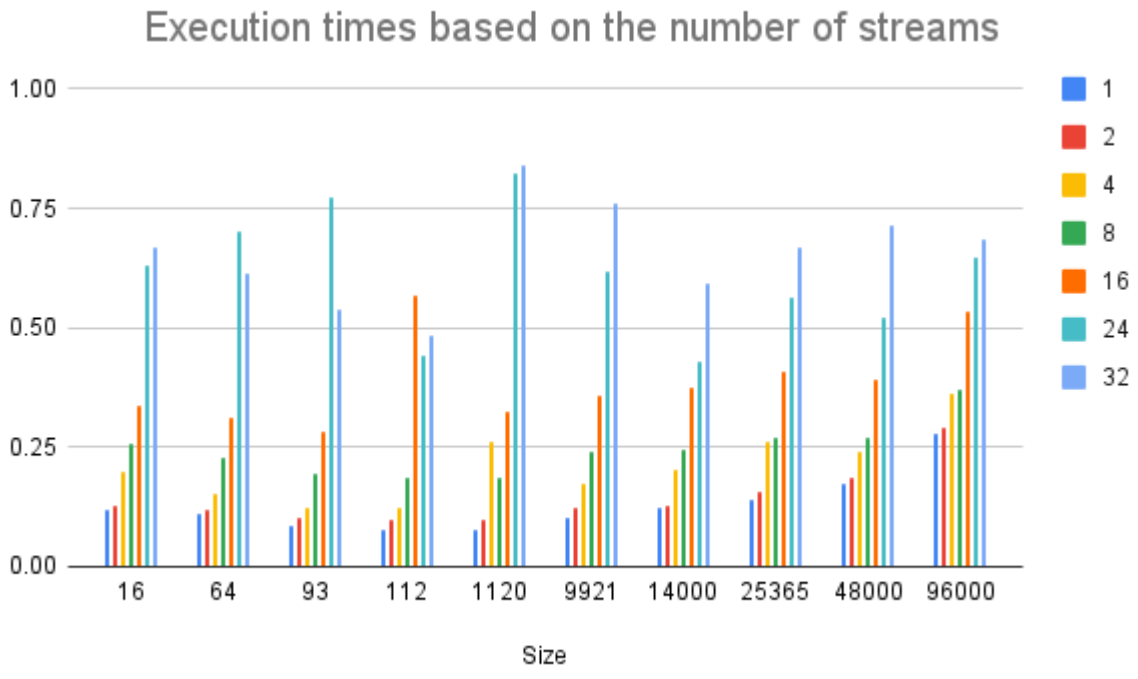
Therefore, one should consider using pinned memory when there is a lot of data transfer between GPU device and the host.

Also, as the pinned memory stays on the host memory, to support async memory operations, it is often required to use pinned memory.

Pinned memory does not require an implicit copy from host memory, which saves the host memory bandwidth and enables faster data transfer between host and GPU.

Q4 & Q5

System information: CSE-EDU cluster with srun command (titanxp) with 256 block size.



Raw data:

Size	1	2	4	8	16	24	32
16	0.119752	0.128109	0.198017	0.257453	0.337547	0.632046	0.669443
64	0.112953	0.118043	0.152927	0.226408	0.314027	0.703569	0.613272
93	0.084853	0.100635	0.123166	0.193246	0.281561	0.771914	0.539739
112	0.078705	0.099164	0.125001	0.187266	0.568077	0.443868	0.48545
1120	0.07891	0.097602	0.2608	0.188069	0.325492	0.821562	0.84
9921	0.100721	0.123453	0.173905	0.242378	0.359857	0.61922	0.760629
14000	0.121778	0.128389	0.201847	0.246471	0.376275	0.430965	0.593943
25365	0.140144	0.156506	0.259839	0.269761	0.410542	0.565409	0.669685
48000	0.174405	0.186103	0.239172	0.269613	0.389886	0.519939	0.716519
96000	0.280555	0.29057	0.361885	0.372482	0.532907	0.648405	0.685629